DOCUMENT RESUME

ED 219 413 TM 820 424

AUTHOR Myette, Beverly M.; White, Karl R.
TITLE Selecting an Appropriate Multiple Comparison Technique: An Integration of Monte Carlo Studies.
PUB DATE Mar 82
NOTE 22p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Hypothesis Testing; Sampling; *Statistical Analysis; *Statistical Bias; Statistics
IDENTIFIERS *Monte Carlo Studies; *Multiple Comparisons; T Test; Tukey Statistic; Type I Errors

ABSTRACT

Twenty Monte Carlo studies on multiple comparison (MC) techniques were conducted to examine which MC technique was the "method of choice." The results from these studies had several apparent contradictions when different techniques were investigated under varying sample size and variance conditions. Box's coefficient of variance variation and bias coefficient were used to quantify the degree of variance heterogeneity and estimated bias of the Type I error rate for the sample and variance condition with each technique studied. Twenty-five multiple comparison techniques were covaried with study characteristics. Wide disparity in data limited recommendation of a "method of choice." Results indicated that the Kramer modification for unequal sample sizes was the preferred method when the Tukey MC technique was elected by the researcher. Neither of Box's coefficients were found to predict bias in Type I error rate. A simple t test administered after a significant omnibus F test (the "Two-stage t-test") has been obtained, appeared to be the most simple and accurate procedure for multiple comparisons. Further replications of this technique are needed to confirm its accuracy. (Author/DWH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Abstract

Monte Carlo studies on multiple comparison (MC) techniques have been conducted to examine which MC technique is the "method of choice." However, the results from these studies have a number of apparent contradictions when different techniques are investigated under varying sample size and variance conditions. Box's coefficient of variance variation and bias coefficient were used to quantify the degree of variance heterogeneity and estimated bias of the Type I error rate for the sample and variance condition with each technique studied (n = 1579). Twenty-five multiple comparison techniques were covaried with study characteristics. Wide disparity in data limited recommendation of a "method of choice." However, results indicated that once the Tukey MC technique was elected by the researcher, the Kramer modification for unequal sample sizes is the preferred method. Furthermore, neither of Box's coefficients were found to predict bias in Type I error rate. Finally, a simple *t* test administered only after a significant omnibus F test has been obtained appears to be the most parsimonious and accurate procedure for multiple comparisons. However, only two Monte Carlo studies have examined this technique, therefore, further replication is needed.

ED219413

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

- This document has been reproduced as received from the person or organization originating it.
- Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Selecting an Appropriate Multiple Comparison Technique: An Integration of Monte Carlo Studies

Beverly M. Myette

and

Karl R. White, Ph.D.

Utah State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B.M. Myette

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the Annual Meeting of the American Educational Research Association in New York, March 1982

TM 820 424

Selecting An Appropriate Multiple Comparison Technique:

An Integration of Monte Carlo Studies

Introduction

"The experimenter seeking methods to compare several groups suffers from an embarrassment of riches seldom found in statistics" (Petrinovich & Hardyck, 1969, p. 53).

Comparison of treatment means in educational research frequently involves the use of multiple comparison (MC) techniques. If an omnibus F test results in statistically significant differences for an independent variable with more than two levels, the researcher must select from an ever increasing array of available MC techniques. The selection process is complex and not without confusion.

A number of authors (e.g., Begun & Gabriel, 1981; Swaminathan & DeFriesse, 1979; Keselman & Rogan, 1977; Johnson & Jones, 1972; Games, 1971; Mehta & Srinivasan, 1970; Aitkin, 1969; Ryan, 1959) have described theoretically, mathematically, and preferentially various multiple comparison techniques from which the researcher can choose. One excellent paper by Hopkins and Anderson (1974) proposed a logical paradigm which outlined the factors to be considered in selecting the "method of choice." Although useful, this paradigm approaches the question of selecting the most appropriate MC technique for a particular situation only from a logical and theoretical perspective. Because the data were largely unavailable at the time that the Hopkins and Anderson (1974) paper was written, another body of literature was not considered which could shed substantial additional light on the question of which MC technique should be the method of choice for a particular situation. These data come from 20 different Monte Carlo studies which have empirically considered the accuracy of various MC techniques in

terms of Type I and Type II error rates. Of these Monte Carlo studies, 14 were done after the Hopkins and Anderson (1974) article.

In spite of the potential value of these Monte Carlo studies, they have not always produced easily understandable or consistent results. Based on a Monte Carlo study which considered the actual versus predicted Type I error rates of 10 different MC techniques, Carmer and Swanson (1974) concluded that using the Duncan Multiple Range Test resulted in fewer Type I errors than predicted. However, Petrinovich and Hardyck (1969) conducted a Monte Carlo study from which they concluded that the Duncan Multiple Range test resulted in significantly more Type I errors than predicted. Similarly, Petrinovich and Hardyck (1969) concluded that Scheffé and Tukey procedures both resulted in unacceptably high levels of Type II error rates whereas Keselman and Toothaker (1973) concluded that both the Scheffé and Tukey procedures were quite sensitive to detecting true differences (i.e., had low Type II error rates).

Although the results of existing Monte Carlo studies provide a potentially rich data source for understanding when it is more appropriate to use which multiple comparison technique, there has not been a systematic integration of the results from the Monte Carlo studies. The apparent contradictions in the existing studies underscore the need for such a systematic analysis.

Rationale and Procedures

The quantitative integration or meta-analysis techniques proposed by Glass (1976) provide a promising methodology for making sense out of the literature on multiple comparison techniques. The data source for the integration consisted of 20 Monte Carlo studies listed in the references. For each MC technique considered, outcome information was quantitatively

coded in terms of deviation from expected Type I error rates in two ways:

$$\text{Effect Size \#1 (ES1)} = \frac{\text{obtained Type I error rate}}{\text{nominal Type I error rate}} \quad (1)$$

$$\text{Effect Size \#2 (ES2)} = \text{obtained Type I error rate} - \text{nominal Type I error rate} \quad (2)$$

Although the two measures of effect size are highly correlated for a given set of data, they do give qualitatively different estimates of the degree of error. For example, Table 1 shows calculations for ES1 and ES2 for a small set of contrived data. As can be seen in Table 1, the correlations between ES1 and ES2 with the same nominal Type I error rate level will always be 1.00. However, when different nominal Type I error rates for ES1 and ES2 are combined, the correlation is less than 1.00 although still fairly high. Because the two different measures of Effect Size yield logical, but different estimates of accuracy (in terms of Type I error rates), it is important to consider the results in both ways.

The covariation of these outcomes with variables such as type of Monte Carlo procedures used, sample size, number of groups, degree of heterogeneity among samples, and sample characteristics across the 20 Monte Carlo studies could be investigated to try and resolve the contradictory conclusions of previous studies.

One such sample characteristic which seemed particularly relevant for MC situations with unequal sample sizes and unequal variances was C^2, an index of the degree to which the variances for different groups are heterogeneous. Although many of the Monte Carlo studies have included samples with supposedly heterogeneous variance in their models, Rogan, Keselman, and Breen

Table 1
Contrived Data Demonstrating Differences
Between ES1 and ES2

	Nominal	Obtained	ES1	ES2
1.	.01	.015	1.5	.005
2.	.01	.02	2.0	.01
3.	.01	.005	.5	-.005
4.	.01	.001	.1	-.009
5.	.01	.012	1.2	.002
6.	.01	.018	1.8	.008
7.	.01	.008	.8	-.002
8.	.01	.003	.3	-.007
9.	.01	.002	.2	-.008
10.	.01	.01	1.0	0
11.	.05	.07	1.4	.02
12.	.05	.08	1.6	.03
13.	.05	.03	.6	-.02
14.	.05	.02	.4	-.03
15.	.05	.015	.3	-.035
16.	.05	.05	1.0	0.0
17.	.05	.06	1.2	.01
18.	.05	.10	2.0	.05
19.	.05	.04	.8	-.01
20.	.05	.01	.2	-.04

$$\Pi_{ES1 \cdot ES2_{.01}} = 1.0$$

$$\Pi_{ES1 \cdot ES2_{.5}} = 1.0$$

$$\Pi_{ES1 \cdot ES2_{.01\ \&\ .05}} = .81$$

(1977) suggested that conclusions about the presence or absence of heterogeneous variances between groups are frequently incorrect. They proposed the Box coefficient of variance variation (C^2) as a suitable technique for defining the degree of variance heterogeneity for a given sample. According to Box (1954), C^2 is defined as:

$$C^2 = \frac{\left[(N-K)^{-1} \sum_{k}^{K} \nu_k (\sigma_k^2 - \dot{\sigma}^2)^2\right]^{\frac{1}{2}}}{\dot{\sigma}^2} \qquad (3)$$

where $\dot{\sigma}^2 = \frac{\sum_{k}^{K} \nu_k \sigma_k^2}{\sum V_k}$

and $\nu_k = n_k - 1$

In addition to the C^2 coefficient, Rogan et al. (1977) recommended the Box bias coefficient (b) to estimate the discrepancy that would be expected between the actual and nominal probabilities of a Type I error under conditions of heterogeneous variances. According to Box (1954), the bias coefficient b is defined as:

$$b = 1 + \frac{1 - 1/N}{1 - 1/K} \left[\frac{\bar{\sigma}^2}{\dot{\sigma}^2} - 1\right] \qquad (4)$$

Many of the existing Monte Carlo studies have considered cases of unequal variances and unequal n's, but few authors have used Box's coefficients. Rogan et al. (1977) were the only authors to recommend consistent use of C^2 as an estimate of the degree of variance heterogeneity. The importance of C^2 in trying to interpret the results of previous Monte Carlo studies on MC techniques can be seen by the following data from Rogan et al. (1977).

	Case A	Case B
n_k :	12, 16, 18, 20, 24, 30	6, 12, 14, 16, 24, 48
σ^2 :	.1, .4, .7, 1.0, 1.2, 2.6	.642, .642, .642, .642, .642, 2.79
ratio σ^2 :	1:4:7:10:12:26	1:1:1:1:1:4.35
C^2 :	.696	.693

As can be seen, even though the two cases exhibit very different ratios of unequal variances, they are virtually identical with respect to the degree of variance heterogeneity as defined by the coefficient of variance variation. Consequently, even though previous Monte Carlo studies may have used different ratios of variances to examine the effect of unequal variances, the effect of different degrees of variance heterogeneity on Type I error rates may not have been investigated.

Examining how C^2 and b covaried with the accuracy of Type I error rates using different MC techniques had two potential benefits:

1. C^2 and b provide consistent techniques for defining the degree of variance heterogeneity and predicted bias which might explain in part why previous Monte Carlo studies have reached contradictory conclusions.
2. If b is a good predictor of Type I error rate, it would be very useful to researchers using inferential statistics with samples of unequal variances and unequal sample sizes.

To obtain the studies to be considerd in the meta-analysis, a computer search, bibliographic search, and hand search were conducted to find Monte Carlo studies which had examined the relative accuracy of two or more MC techniques. Twenty studies were collected. Study characteristics included for coding were: type of multiple comparison technique (including type of correction for unequal sample sizes), alpha base error rate (i.e., pairwise

or experimentwise procedures), number of groups, sample size and variance conditions, replications, computer program, pseudo-random number generator procedure, and nominal and obtained Type I error rate. Box's coefficients, C^2 and b, were computed after all studies were coded.

Results and Discussion

Classical Meta-Analysis

The classical approach to meta-analysis requires the outcome of each study to be coded in a common metric (Effect Size) and the covariance of different study characteristics with Effect Size to determine if study characteristics are associated with different outcomes. To do this, the data are broken down into different one-, two- and three-way combinations to examine how the outcomes vary for different values of the study characteristics. For example, the Effect Size from this study (accuracy of Type I error rate) could be examined for each technique broken down by pairwise and experimentwise base error rates, and different combinations of equal/unequal sample size and equal/unequal variances (1) $=n/=s^2$; (2) $\neq n/\neq s^2$; (3) $\neq n/=s^2$; (4) $=n/\neq s^2$). As can be seen, such a breakdown would result in a matrix with 200 cells (25 techniques X 2 base error rates X 4 n/s^2 combinations). Unless sufficient studies are available for coding, the interpretation of such breakdowns becomes difficult because of empty cells and cells with only a few effect sizes.

Unfortunately, from the 20 studies collected, only 12 reported sufficient information about sample size and variance conditions so they could be coded (1,579 Effect Sizes). Within these studies, 25 different multiple comparison techniques were investigated. Most of these techniques were examined under limited sample and variance conditions. Consequently, the classical approach to meta-analysis was not very useful in quantitatively

integrating the Monte Carlo studies and attempting to address the most important research questions.

Table 2 illustrates the myriad of multiple comparison techniques investigated for both the codable and uncodable studies. Most authors studied a number of MC techniques, but frequently (approximately 50% of the time) many techniques were only considered once or twice. Of the 27 techniques, almost half (16) were Tukey modifications. Furthermore, only half (7) of the Tukey modifications were recommended by the authors. Columns 4 and 5 of Table 2 show the ES1 for those techniques included in the meta-analysis for which more than 20 ES's could be coded. As can be seen, only nine techniques were examined frequently enough to even be included in the table, and seven of these were a modification of Tukey, or a procedure for dealing with unequal n's when using Tukey. Also, there were no cells in Table 2 for the unequal sample and unequal variance conditions with enough ES's to be included in the table. Even though a total of 265 ES's were calculated for these conditions ($\neq n/\neq s^2$), the distribution across techniques and alpha base error rate was such that no cell had more than 20 ES's.

The Kramer procedure for using the Tukey technique for samples with unequal n's was the most frequently examined and the most frequently recommended technique. As can be seen in Table 2, Kramer's method had the smallest deviation from alpha and was slightly conservative. The Scheffé technique was also frequently recommended and had an extremely conservative average ES. Not surprisingly, the LSD experimentwise error was high and rejected by all three authors who examined it. Similar results were obtained when deviation from nominal Type I error rate was examined using ES2. These results are reported in Table 3.

Table 2
Breakdown of the 27 Techniques Considered in the Monte Carlo Studies

	(1) No. of Times Examined	(2) No. of Times Recommended	(3) No. of Times Rejected	(4) ES1 Pairwise	(5) ES1 Experimentwise	(6) Source
Scheffe	5	4		21 $n=22$ $(=n,=s^2)$	.30 $n=31$ $(=n,=s^2)$	16*[c],1*,3*,2*,12,11
					1.09 $n=53$ $(=n,=s^2)$	
Tukey (WSD)	7	3		2.36 $n=30$ $(=n,\neq s^2)$	1.20 $n=25$ $(=n,=s^2)$	16*,1*,3*,2*,17*,14,9
Winer[d]	7	2	1		1.42 $n=248$[b] $(\neq n,=s^2)$	8*,19*,10*,18,13,12,4
Kramer[d]	8	5			.93 $n=261$[b] $(\neq n,=s^2)$	5*,19*,10*,18,13,11,4,7
Miller[d]	4				1.70 $n=74$[b] $(\neq n,=s^2)$	19*,17*,18,13
Newman-Keuls	5					16*,1*,3*,2*,17*
Duncan	4	1	1			16*,1*,3*,2*
POC	1					15*
LSD	3		3	1.05 $n=22$ $(=n,=s^2)$	5.05 $n=30$ $(=n,=s^2)$	16*,3*,2*
2-stage t-test	2	2				1*,3*
Behrens-Fisher	2	1				14,7
Dalal Modif.[d]	1					20*
Games & Howell[d]	2	2		.92 $n=30$ $(=n,\neq s^2)$		20*,11
Games & Howell Modif.[d]	1					5*
Tamhane 2[d]	2	1		.65 $n=26$ $(=n,\neq s^2)$		20*,5*
Tamhane 2 Modif.[d]	1					5*
Spjotvoll & Stoline[d]	4					20*,11,4,6
Brown & Forsythe[d]	2	1				20*,9
Hochberg GT2[d]	3					20*,4,6
Cochran[d]	1	1		.71 $n=29$ $(=n,\neq s^2)$		5*
Spjotvoll Modif.[d]	1					20*
Hochberg H1[d]	3					20*,11,9
Bayesian BSD	1	1				3*
Bayesian BED	1	1				3*
Gabriel	1					4
Geneize & Hochberg[d]	1					4
Tukey with Welch df Modif[d]	1	1				9

[a]Only ES1 reported and those data where n>20 in meta-analysis for the different sample and variance conditions.
[b]Smith, 1968 reference.
[c]Codable articles asterisked.
[d]Tukey modification.

Table 3
ES2 for MC Techniques with n>20

	Pairwise	Experimentwise
Scheffé	-.039 n = 24 (=n, $=s^2$)	-.035 n = 31 (=n, $=s^2$)
Tukey	.068 n = 30 (=n, $\neq s^2$)	.004 n = 53 (=n, $=s^2$) .010 n = 25 (=n, $\neq s^2$)
Winer		.006 n = 248 ($\neq$n, $=s^2$)
Kramer		-.003 n = 261 ($\neq$n, $=s^2$)
Miller		.013 n = 74 ($\neq$n, $=s^2$)
LSD	.0002 n = 22 (=n, $=s^2$)	.202 n = 30 (=n, $=s^2$)
Games & Howell	-.004 n = 30 (=n, $\neq s^2$)	
Tamhane	-.017 n = 26 (=n, $\neq s^2$)	
Cochran	-.015 n = 29 (=n, $\neq s^2$)	

In summary, even though researchers have investigated a variety of multiple comparison techniques resulting in 1,579 Effect Sizes, the data are spread so thinly over so many different techniques that no clear conclusions about the method of choice can be reached using the meta-analysis approach. If consumers and reviewers are to draw viable conclusions from these Monte Carlo studies, additional data must be reported in descriptions of previous studies. Unfortunately, this frequently has not been done.

Even though some of the most important questions concerning the selection of MC technique could not be examined, the data do provide some information on several important questions:

1. Multiple comparisons with unequal sample sizes. Even though the data from previous Monte Carlo studies do not suggest the MC "method of choice," they do provide guidance about which procedure should be

used with samples having unequal n's if the Tukey MC technique has already been selected by the researcher. That is, once the researcher has decided to use the Tukey multiple comparison technique, the Kramer method for unequal sample sizes is most accurate in terms of agreement between nominal and obtained Type I error rate. Kramer's technique uses the harmonic mean from the two groups being compared. Unfortunately, as shown in Table 2, the data from which this conclusion is drawn came from only one study. Given the contradictions between other Monte Carlo studies regarding the single question of choice, it is important that these data be replicated using the same as well as different n/s^2 conditions.

2. Coefficients of variance variation and bias. As noted earlier, both C^2 and b were recommended as variables which might be important in explaining why previous Monte Carlo studies had reached contradictory conclusions. Even though the codable Monte Carlo studies did not provide sufficient data to answer this question, other important information about C^2 and b was obtained.

For each Effect Size obtained, both C^2 and b were calculated. The correlation between ES and C^2 or b provides information about whether either are good at predicting the accuracy of Type I error rates. Of course, the C^2 and b coefficients only make sense when used with unequal variances and unequal sample sizes. With equal sample sizes, regardless of the magnitude of the unequal variances, C^2 will always equal zero and b equal 1.00.

Table 4 shows the correlation of ES1 with both C^2 and b for unequal variances and unequal sample sizes (all other correlation coefficients were zero or less than n = 10).

Table 4
Pearson Product Moment Correlations of ES1 with C^2 and b
for Monte Carlo Studies Examining Samples with
Unequal n and Unequal Variances[a]

	Pairwise		Experimentwise	
	C^{2a}	b[a]	C^2	b
OVERALL	.136 n = 98	-.309 n = 98	.346 n = 179	.086 n = 179
Tukey	.249 n = 19	-.750 n = 19	.330 n = 27	.043 n = 27
Kramer	.247 n = 17	-.761 n = 17	.060 n = 12	-.172 n = 12
POC			.187 n = 12	.213 n = 12
Behrens Fisher			.020 n = 12	-.212 n = 12
Games & Howell	.675 n = 15	-.261 n = 15		
Tamhane 2	.617 n = 19	-.244 n = 19		
Tamhane 2 Modification	.548 n = 15	-.240 n = 15		

$R^2_{ES1 \cdot C^2} = .00065$

$R^2_{ES1 \cdot b} = .00038$

[a]Only ES1 with n>10 included.

As can be seen, the amount of variance in ES1 accounted for by either C^2 or b was generally very low. Even more disconcerting is the instability of the correlations from technique to technique. The only cells where the correlations were moderately strong (Tukey & Tukey/Kramer pairwise) were based on relatively small numbers of ES's. The data in Table 4 do not support the use of Box's b coefficient as a means of estimating the bias in the Type I error rate.

3. t test in combination with significant omnibus F. In recent years there has been a proliferation of multiple comparison techniques and modifications (see Table 2). In spite of the 20 Monte Carlo studies which have examined the Type I error rates of these techniques, researchers do not agree as to the best MC method for different situations. However, two of the Monte Carlo studies raise an important possibility that needs further investigation. Carmer and Swanson (1973) and Bernhardson (1975) investigated using a simple t test which would only be used if preceded by a significant omnibus F test as a multiple comparison technique. This "two-stage t test" should not be confused with the Least Significant Difference (LSD) multiple t test which does not require a significant F test to be used.

Both of the studies that investigated the accuracy of the "two-stage t test" found this technique to control pairwise and experimentwise Type I error rates. One of the major criticisms against using multiple t tests has been the high Type I experimentwise error rate (particularly as the number of treatment means increases). Carmer and Swanson (1973) and Bernhardson (1975) did not find this to be the case with this "two-stage t test." With K = 10, Bernhardson (1975) found the true Type I experimentwise error rate to be .059. Similarly, with K = 20, Carmer and Swanson (1973) found the true Type I experimentwise error rate in their Monte Carlo study to be .051. These authors recommended the "two-stage t test" as a simple and accurate multiple comparison technique.

Given the almost bewildering array of multiple comparison techniques now available to the researcher, there is much to recommend such a parsimonious solution to the problem of selecting a

MC technique. Based on existing data, the simple t test preceded by a significant F test appears to be as accurate a technique as any--and much simpler to use. Consequently, the "two-stage t test" certainly warrants consideration and further investigation.

Summary and Recommendations

Twenty Monte Carlo studies which examined the accuracy of obtained Type I error rates for different multiple comparison techniques were examined. Although there was not enough codable data available to conduct a classical meta-analysis, information was provided on several important research questions.

1. When the researcher has decided to use the Tukey multiple comparison technique, the Kramer method appears to be the best method to use with samples having unequal sizes and equal variance situations.
2. Box's coefficient of variance variation and bias coefficient were not found to be related to Type I error rate. At this point, Box's b coefficient should not be used to estimate Type I error bias.
3. The simple t test preceded by a significant F test (the "two-stage t test") warrants serious consideration as an accurate and simple MC technique. However, further replications of the Carmer and Swanson (1973) and Bernhardson (1975) studies need to be conducted. If further replications confirm that the "two-stage t test" is as accurate as it appears to be, then the extensive work in developing new techniques and modifications of existing techniques may be focused in the wrong areas. Instead of creating more techniques, it is more important to systematically integrate the information we now have and to determine if this simple approach is not only more parsimonious, but just as accurate.

References

Aitkin, M. Multiple comparisons in psychological experiments. The British Journal of Mathematical and Statistical Psychology, 1969, 22(2), 193-198.

Begun, J., & Gabriel, K. R. Closure of the Newman-Keuls multiple comparisons procedure. Journal of the American Statistical Association, 1981, 76(374), 241-245.

Bernhardson, C. Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. Biometrics, 1975, 31, 229-232.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Annals of Mathematical Statistics, 1954, 25, 290-302.

Carmer, S., & Swanson, M. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 1973, 68(341), 66-74.

Games, P. A. Multiple comparisons of means. American Educational Research Journal, 1971, 8(3), 531-565.

Glass, G. V. Primary, secondary and meta-analysis of research. Educational Researcher, 1976, 5, 3-8.

Glass, G. V, McGaw, B., & Smith, M. L. Integration of research studies: Meta-analysis of social research. Beverly Hills, CA: Sage Publishing Co., 1981.

Hopkins, K., & Anderson, B. A guide for selecting the "method of choice." Laboratory of Educational Research, University of Colorado, 1974.

Johnson, R., & Jones, L. Multiple comparisons and error rate. Journal of College Student Personnel, 1972, 13(2), 154-158.

Keselman, H., & Rogan, J. The Tukey multiple comparison tests: 1953-1976. Psychological Bulletin, 1977, 84(5), 1050-1056.

Keselman, H., & Toothaker, L. Error rates for multiple comparison methods: Some evidence concerning the misleading conclusions of Petrinovich and Hardyck. Psychological Bulletin, 1973, 80(1), 31-32.

McGaw, B., & White, K. R. Meta-analysis of empirical research. Training materials for AERA Research Training Presession, April 1981.

Mehta, J., & Srinivasan, R. On the Behrens-Fisher problem. Biometrics, 1970, 57, 649-655.

Miller, R. G., Jr. Developments in multiple comparisons: 1966-1976. Journal of the American Statistical Association, 1977, 72(360), 779-788.

Petrinovich, L., & Hardyck, C. Error rates for multiple comparison methods: Some methods concerning the frequency of erroneous conclusions. Psychological Bulletin, 1969, 71, 43-54.

Rogan, J., Keselman, H., & Breen, L. Assumption violations and rates of Type I error for the Tukey multiple comparison test: A review and empirical investigation via a coefficient of variance variation. Journal of Experimental Education, 1977, 46, 20-25.

Ryan, T. Multiple comparisons in psychological research. Psychological Bulletin, 1959, 56(1), 26-47.

Swaminathan, H., & DeFriesse, F. Detecting significant contrasts in analysis of variance. Educational and Psychological Measurement, 1979, 39, 39-44.

References (Monte Carlo Studies)

*1. Bernhardson, C. Type I error rate when multiple comparison procedures follow a significant F test of ANOVA. Biometrics, 1975, 31, 229-232.

*2. Boardman, T. J., & Moffitt, D. R. Graphical Monte Carlo Type I error rates for multiple comparison procedures. Biometrics, 1971, 738-744.

*3. Carmer, S. G., & Swanson, M. R. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 1973, 68(341), 66-74.

4. Dunnett, C. W. Pairwise multiple comparisons in the homogeneous variance, unequal sample size case. Journal of the American Statistical Association, 1980, 75(372), 789-795.

*5. Dunnett, C. W. Pairwise multiple comparisons in the unequal variance case. Journal of the American Statistical Association, 1980, 75( ), 796-800.

6. Gabriel, K. R. A simple method of multiple comparisons of means. Journal of the American Statistical Association, 1978, 73( ), 724-729.

*7. Games, P. A., & Howell, J. F. Pairwise multiple comparison procedures with unequal n's and/or variances: A Monte Carlo study. Journal of Educational Statistics, 1976, 1(2), 113-125.

*8. Howell, J. F., & Games, P. A. The robustness of the analysis of variance and the Tukey WSD test under various patterns of heterogeneous variances. The Journal of Experimental Education, 1973, 41(4), 33-37.

9. Keselman, H. J., Games, P., & Rogan, J. Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus test statistic. Psychological Bulletin, 1979, 86(4), 884-888.

*10. Keselman, H. J., Murray, R., & Rogan, J. Effect of very unequal group sizes on Tukey's multiple comparison test. Educational and Psychological Measurement, 1976, 36, 263-270.

11. Keselman, H. J., & Rogan, J. A comparison of the modified Tukey and Scheffe methods of multiple comparisons for pairwise contrasts. Journal of the American Statistical Association, 1978, 73(361), 47-52.

12. Keselman, H. J., & Toothaker, L. E. Comparison of Tukey's T-method and Scheffe's S-method for various numbers of all possible differences of average contrasts under violations of assumptions. Educational and Psychological Measurement, 1974, __, 511-519.

13. Keselman, H. J., Toothaker, L. E., & Shooter, M. An evaluation of two unequal n_k forms of the Tukey multiple comparison statistic. Journal of the American Statistical Association, 1975, 70(351), 584-587.

14. Kohr, R. L., & Games, P. A. Testing complex a priori contrasts on means from independent samples. Journal of Educational Statistics, 1977, 2(3), 207-216.

*15. Levy, K. A priori contrasts under conditions of variance heterogeneity. Journal of Experimental Education, 1978, 42-45.

*16. Petrinovich, L. F., & Hardyck, C. D. Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 1969, 71, 43-54.

*17. Ramseyer, G. C., & Tcheng, T. The robustness of the studentized range statistic to violations of the normality and homogeneity of variance assumptions. American Educational Research Journal, 1973, 10(3), 235-240.

18. Rogan, J. C., Keselman, H. J., & Breen, L. J. Assumption violations and rates of Type I error for the Tukey multiple comparison test: A review and empirical investigation via a coefficient of variance variation. Journal of Experimental Education, 1977, 20-25.

*19. Smith, R. A. An empirical analysis of the effect of unequal sample size on the Tukey studentized range technique. University of Colorado, Laboratory of Educational Research, No. 18, 1968.

*20. Tamhane, A. C. A comparison of procedures for multiple comparisons of means with unequal variances. Journal of the American Statistical Association, 1979, 74(366), 471-480.

---

*Codable Monte Carlo studies.